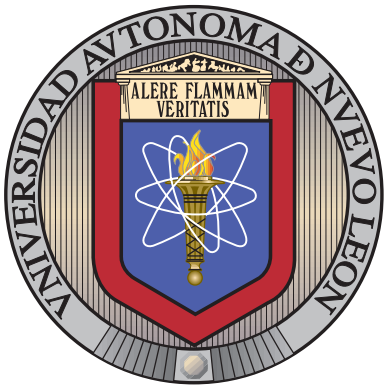
Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

Subdirección Académica



Modelado y visualización de relaciones entre contaminantes del aire y salud pública

por

Selene Berenice Prado Prado

como requisito parcial para obtener el grado de

ingeniería en tecnología de software

Julio 2022

Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

Subdirección Académica

Modelado y visualización de relaciones entre contaminantes del aire y salud pública

por

Selene Berenice Prado Prado

como requisito parcial para obtener el grado de

ingeniería en tecnología de software

Julio 2022

Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

Subdirección Académica

Los miembros del Comité de Tesis recomendamos que la Tesis «Modelado y visualización de relaciones entre contaminantes del aire y salud pública», realizada por el alumno Selene Berenice Prado Prado, con número de matrícula 1810042, sea aceptada para su defensa como requisito parcial para obtener el grado de Ingeniería en Tecnología de Software.

El Comité de Tesis

Dra. Satu Elisa Schaeffer	Dra. Sara Elena Garza Villarreal
Coasesora	Coasesora
Dr. José Arturo Berrones Santos	Dr. Romeo Sánchez Nigenda
Revisor	Revisor

Vo. Bo.

Dr. Fernando Banda Muñoz

Subdirección Académica

San Nicolás de los Garza, Nuevo León, julio 2022

Índice general

Índice de figuras

Índice de cuadros

Agradecimientos

Quiero agradecer a la Dra. Elisa por el apoyo durante el desarrollo de mi tesis y por la motivación y conocimientos brindados para seguir desarrollándome profesionalmente en lo que me gusta. Al programa PAICYT-UANL por su contribución brindada bajo las claves CE1421-20 y CE1842-21.

A mis padres, Lilia Prado López y Adan Alfaro Lerma, por su apoyo y motivación constante desde siempre. A mis hermanos Angel, Estrella, y Adali, a quienes he visto crecer y de quienes he aprendido mucho.

Resumen

Selene Berenice Prado Prado.

Candidato para obtener el grado de Ingeniería en Tecnología de Software.

Universidad Autónoma de Nuevo León.

Facultad de Ingeniería Mecánica y Eléctrica.

Título del estudio: Modelado y visualización de relaciones entre contaminantes del aire y salud pública.

Número de páginas: 52.

Objetivos y método de estudio: El objetivo de la investigación es generar modelos que permitan visualizar relaciones entre contaminantes atmosféricos y salud pública. Los modelos generados se utilizan en conjunto con datos obtenidos de la Secretaría de Salud del Gobierno de México y registros de los niveles de los contaminantes presentes en el área metropolitana de Monterrey.

El tener un modelo que permita visualizar relaciones entre contaminantes atmosféricos y salud pública que sea utilizado con datos confiables y verídicos pueden ayudar a entender el impacto que tiene el aumento del nivel de contaminantes atmosféricos.

Contribuciones y conclusiones: En la investigación se exploran diversas maneras de visualizar la información, además de generar modelos que permiten analizar las relaciones existentes entre los niveles de determinados contaminantes y CIE. Las visualizaciones generadas son series de tiempo y gráficos de radar, y modelos de regresión lineal y de regresión lineal múltiple.

Firma de la coasesora: _____

Dra. Satu Elisa Schaeffer

Firma de la coasesora: _____

Dra. Sara Elena Garza Villarreal

Capítulo 1

Introducción

El *aprendizaje máquina*[1] es un área dentro de la *ciencia de datos*[2] que puede ayudar a crear dichos modelos para tener una más eficiente visualización cuando se trabaja con una gran cantidad de datos, que es lo que se requiere para el presente trabajo. El área de la ciencia de datos es muy útil ya que permite trabajar con grandes cantidades de datos aminorando la cantidad de tiempo empleado en la creación de gráficos que permitan visualizar los datos. El crear modelos para la visualización de datos ayuda a observar con mayor claridad los datos para encontrar relaciones entre ellos.

La tarea en el presente proyecto es utilizar modelos para visualizar las relaciones entre los contaminantes del aire y salud pública. Para la realización de los experimentos se tienen datos de ingresos hospitalarios provenientes de la base de datos de la Secretaría de Salud del Gobierno de México [9]. También se tienen registros de los niveles de algunos contaminantes del aire presentes en el área metropolitana de Monterrey, capturados por las estaciones de monitoreo pertenecientes al Sistema Integral de Monitoreo Ambiental (SIMA) [20] mostradas en la figura 1.1 y en el cuadro 1.1.

[1]Traducido como *machine learning* en inglés, tiene como objetivo desarrollar técnicas que les permitan a las computadoras aprender.

[2]Traducido como *data science* en inglés, involucra métodos para extraer conocimiento de datos, eso con la finalidad de que haya un mejor entendimiento de los datos.

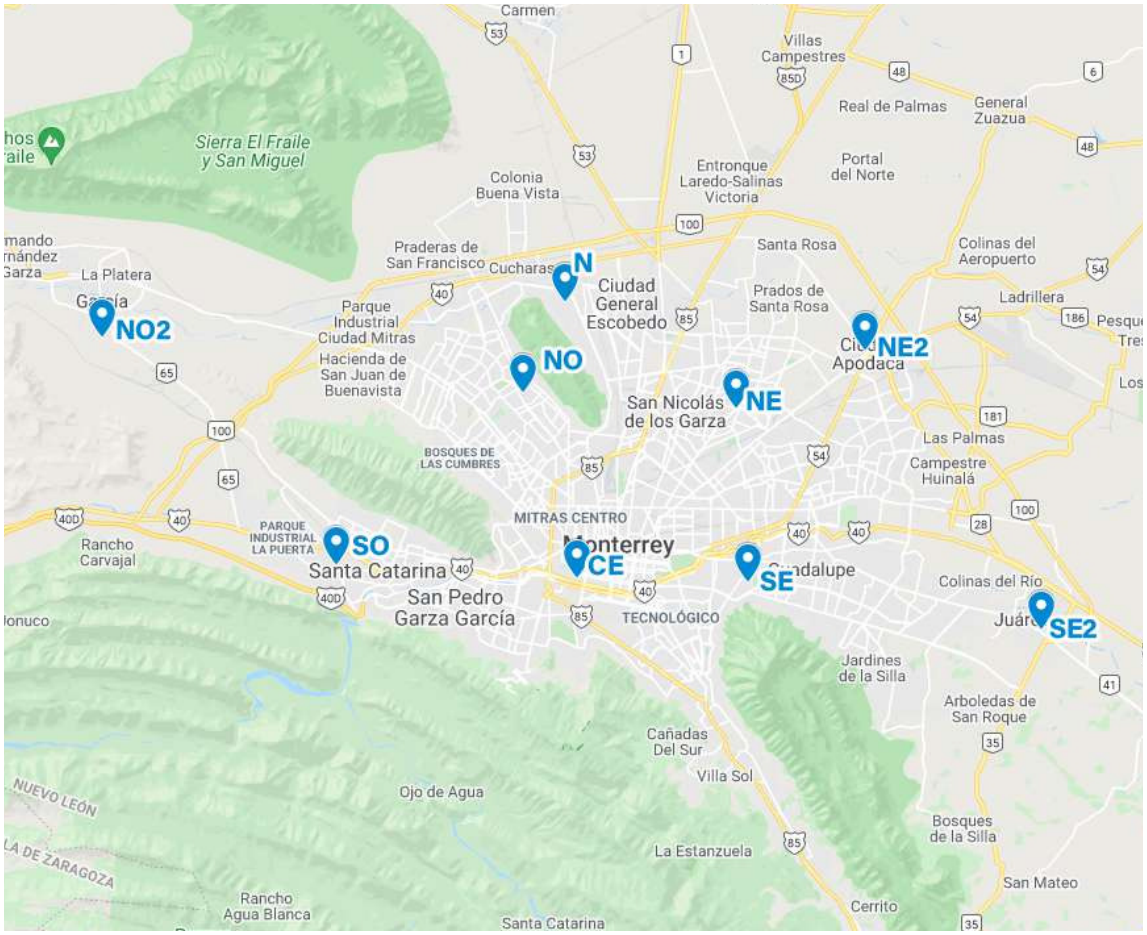


Figura 1.1: Localización de las estaciones de monitoreo de la calidad del aire.

Cuadro 1.1: Coordenadas de las estaciones de monitoreo de la calidad del aire.

Estación	Latitud	Longitud
SE	25.668	-100.249
NE	25.750	-100.255
CE	25.670	-100.338
NO	25.757	-100.366
SO	25.676	-100.464
NO2	25.783	-100.586
NE2	25.777	-100.188
N	25.800	-100.344
SE2	25.646	-100.096

1.1 Motivación

Existen investigaciones que ya han estudiado las relaciones entre contaminantes del aire y salud pública, sin embargo, con el presente trabajo se busca aportar a la creación de nuevas herramientas que permitan observar y estudiar dichas relaciones con el fin de ayudar a tomar medidas adecuadas que permitan aminorar los efectos negativos de los contaminantes del aire en la salud.

1.2 Hipótesis

Los modelos de regresión permiten obtener gráficos donde se pueden observar las relaciones entre el número de ingresos hospitalarios y los niveles de contaminantes del aire.

1.3 Objetivos

En esta sección se establece el objetivo general y los objetivos específicos sobre los que se orienta el presente trabajo.

1.3.1 Objetivo general

Generar, implementar y evaluar modelos que muestran las relaciones existentes entre contaminantes del aire y salud pública tiene la finalidad de apoyar a la implementación de estrategias que aminoran los efectos negativos de los contaminantes del aire en la salud de las personas. Con los modelos generados se puede tener una herramienta que permite identificar gráficamente las relaciones con solo proporcionarle el conjunto de datos.

1.3.2 Objetivos específicos

- Generar, implementar y evaluar modelos de regresión que permiten cuantificar las relaciones entre contaminantes del aire y salud pública a partir de un conjunto de datos.

- Diseñar e implementar visualizaciones interactivas que permiten explorar los modelos implementados y su validez estadística.

- Evaluar la eficacia de los modelos generados para tener noción de la fiabilidad del análisis realizado a partir de los resultados que tales modelos producen.

Capítulo 2

Antecedentes

Existen factores ambientales que afectan la salud de una comunidad como: el abastecimiento de agua potable y el saneamiento, la vivienda y el hábitat, la alimentación, la contaminación ambiental, el empleo de productos químicos y los riesgos ocupacionales [6].

Contaminación del aire es un término usado para describir la presencia de uno o más contaminantes en la atmósfera, cuyas cantidades y características pueden resultar perjudiciales o interferir con la salud, el bienestar u otros procesos ambientales naturales [7].

En este capítulo se presentan los fundamentos y definiciones de los conceptos más relevantes para el tema de estudio abordado.

2.1 Monitoreo de calidad del aire

Existen diversos estudios que muestran la existencia de potenciales efectos a la salud cuando en el aire están presentes contaminantes en forma de partículas, gases o agentes biológicos.

Korc y Sáenz [15] mencionan que desde inicios de 1950 se observa una preocupación por los contaminantes del aire en los países de América Latina y el Caribe. Las universidades y dependencias de los ministerios de salud fueron los organismos que realizaron las primeras mediciones de contaminación en el aire.

En 1965, el Consejo Directivo de la Organización Panamericana de la Salud (OPS) recomendó el establecer programas de investigación de la contaminación del agua y del aire, con el objetivo de colaborar en el desarrollo de políticas adecuadas de control [12].

Mediante el Centro Panamericano de Ingeniería Sanitaria y Ciencias del Ambiente (CEPIS), la OPS acordó establecer una red de estaciones de muestreo de la contaminación del aire. En junio de 1967 la Red Panamericana de Muestreo Normalizado de la Contaminación del Aire (REDPANAIRE) inició sus operaciones recolectando muestras diarias de partículas totales en suspensión (PTS) y de SO_2. La REDPANAIRE comenzó con ocho estaciones y a fines de 1973 tenía un total de 88 estaciones distribuidas en 26 ciudades de 14 países [12].

Para diciembre de 1973 se habían recolectado más de 350,000 datos sobre la calidad del aire, en los que se observa que algunas ciudades mostraban una tendencia al incremento de los niveles de contaminación [12].

En 1980 la REDPANAIRE desapareció y pasó a formar parte del Programa Global de Monitoreo de la Calidad del Aire, iniciado en 1976 por la OMS y el Programa de las Naciones Unidas para el Medio Ambiente (PNUMA), como parte de un sistema global de monitoreo ambiental llamado GEMS por sus siglas en inglés *Global Environmental Monitoring System*.

En la década de 1990, la OMS instituyó, con carácter global, el Sistema de Información para el Control de la Calidad del Aire llamado AMIS por sus siglas en inglés *Air Management Information System*. Entre las actividades más destacadas de AMIS se incluye el coordinar las bases de datos sobre temas relacionados con la calidad del aire.

En Nuevo León, México, las operaciones de la Red Automática de Monitoreo Atmosférico iniciaron en 1993. Dicha red en sus inicios contaba con cinco estaciones fijas de monitoreo continuo de monóxido de carbono (CO), dióxido de azufre (SO_2), óxidos de nitrógeno (NO_x), ozono (O_3) y partículas de tamaño menor a 10 micrómetros (PM10) [15]. Como se muestra en la figura 1.1 y en el cuadro 1.1, actualmente se cuenta con nueve estaciones fijas.

2.2 Series de tiempo

Korc y Sáenz [15] mencionan que las relaciones entre niveles de concentraciones de contaminantes del aire y los efectos sobre la salud generalmente son obtenidas de estudios epidemiológicos de series de tiempo. Uno de los diseños epidemiológicos más utilizados son los estudios de series temporales. Con esos diseños se analizan las variaciones en el tiempo de la exposición al contaminante y el indicador de salud estudiado en una población [2].

Las series de tiempo se pueden definir como un conjunto de observaciones *ot* tomadas en un tiempo t determinado. Los estudios de series de tiempo relacionan estadísticamente los cambios temporales en la repercusión de cambios en la concentración de un contaminante en la población [5].

Para mostrar datos en una serie de tiempo, especialmente en el área médica, estos suelen agruparse en *semanas epidemiológicas*[1]. El agrupamiento en semanas epidemiológicas a diferencia del agrupamiento traducido como *clustering* en inglés que consiste en crear grupos de objetos con similitudes entre ellos a partir de datos no etiquetados, consiste en agrupar los objetos en semanas del año dependiendo de la fecha en que se registró el objeto.

[1]Una semana epidemiológica es un estándar de medición temporal que se utiliza para comparar datos en ventanas de tiempo definidas. La primera semana epidemiológica del año termina el primer sábado de enero de cada año [1].

2.3 Clasificación de enfermedades

Existe un instrumento estadístico y sanitario para identificar enfermedades llamado Clasificación Internacional de Enfermedades (CIE), cuya finalidad es entender las causas de morbilidad y mortalidad de la población y así mejorar la calidad de vida de la misma. Es en base a un criterio epidemiológico y sanitario establecido por Farr a finales del siglo XIX que esta clasificación agrupa enfermedades en epidémicas, generales, locales ordenadas por origen geográfico, trastornos del desarrollo y lesiones [18]. Para lograr distinguirlas se emplea un código alfanumérico que consiste de una letra en la primera posición, seguida de dos dígitos, un punto decimal y un último dígito. El rango de valores va de A00.0 a Z99.9.

2.4 Regresión lineal

La tendencia w_0 de una serie de tiempo puede ser obtenida a partir de una regresión lineal de la misma [8]. Una regresión lineal es una metodología inferencial supervisada que busca predecir valores y dado un vector de variables de entrada t por medio del ajuste de coeficientes w de la función lineal

$$\hat{y}(t, w) = w_0 + w_1 x_1 + ... + w_t x_t. \tag{2.1}$$

2.5 Regresión lineal múltiple

Un modelo de regresión múltiple es un modelo complemento de la regresión lineal simple, el cual tiene dos o más variables independientes k que pueden influir en una variable dependiente y. Peniche-Camps y Cortez-Huerta [19] expresan la regresión múltiple mediante la siguiente ecuación:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon. \tag{2.2}$$

Capítulo 3

Estado del arte

En el presente capítulo se estudia literatura reciente relacionada con el presente trabajo, esto con el objetivo de revisar distintos métodos para resolver el problema planteado en el presente trabajo y, además, también revisar implementaciones similares para resolver problemas distintos. Lo anterior tiene la finalidad de comparar los trabajos revisados e identificar áreas de oportunidad en ellos.

En la primera sección, *trabajos relacionados*, se recopilan obras con características relacionadas al presente trabajo, ya sean relacionados con el problema que se pretende resolver o con los métodos empleados para buscar su resolución.

En la segunda sección, *análisis comparativo*, se comparan las distintas características de los trabajos revisados, de esa forma se pueden determinar las principales ventajas y desventajas de cada trabajo.

Finalmente, en la tercera sección, *áreas de oportunidad*, se realiza una conclusión acerca de los resultados obtenidos del análisis comparativo.

3.1 Trabajos relacionados

Se recopila literatura relacionada desde el año 2017 hasta el año 2021. En esta sección los trabajos se mencionan en orden cronológico tomando en cuenta su año de publicación.

Martín y Bayle [17] estudian la relación entre los niveles de contaminantes ambientales y la presencia de casos de enfermedades respiratorias en las consultas pediátricas. La variable dependiente analizada es la demanda en las consultas pediátricas por bronquiolitis, episodios de broncoespasmo y procesos respiratorios de vías altas. Como variables independientes se tienen los valores de contaminación ambiental. Se calculan coeficientes de correlación y regresión lineal múltiple.

Guarnaccia *et al.* [10] abordan la necesidad de monitoreo, control y predicción de la pendiente de los niveles de contaminantes del aire. Para abordar el problema de investigación utilizan modelos ARIMA.

Julia *et al.* [13] estudian la asociación entre la exposición a largo plazo a la contaminación del aire y la metilación del ADN. Para ello realizan un estudio utilizando modelos de regresión lineal robustos para analizar la asociación entre la exposición al NO_2 y a las partículas PM10 y PM2.5.

Zhang *et al.* [22] en su estudio abordan los niveles de contaminación del aire y su asociación con la presencia de presión sanguínea elevada en niños y adolescentes. La exposición a partículas PM10 y PM2.5 son estimadas con un modelo espacio-temporal. Son utilizados modelos lineales de efectos mixtos y modelos de regresión logística para investigar la asociación entre la exposición a partículas PM y presión sanguínea e hipertensión.

Kim *et al.* [14] estudian la relación entre los niveles de contaminación del aire y la obesidad y problemas cardiometabólicos. Para dicho estudio emplean modelos de regresión lineal.

Liu *et al.* [16] examinan las asociaciones entre la exposición temprana a la contaminación del aire y la incidencia de asma y rinitis alérgica desde el nacimiento hasta la adolescencia. Para su estudio utilizan modelos de regresión.

To *et al.* [21] estudian la asociación entre la exposición temprana a los contaminantes del aire y los egresos hospitalarios por asma. Para su estudio aplican modelos de regresión logística para el análisis de datos.

Breton *et al.* [4] abordan el estudio de la relación entre los niveles de contaminación del aire y el número de admisiones hospitalarias. Para ello se construye un modelo basado en la distribución de Poisson.

Gupta *et al.* [11] estudian la relación entre la mortalidad del coronavirus (COVID-19) y la contaminación del aire. Para dicho estudio emplean un modelo de regresión lineal para establecer la relación entre los parámetros de la contaminación del aire (concentraciones de PM10 o PM2.5) y la variable de respuesta (porcentaje de mortalidad por unidad de casos reportados).

3.2 Comparación de trabajos

La mayoría de los trabajos encontrados emplean modelos de regresión lineal o modelos de predicción. Además, en todos los trabajos encontrados el problema tratado presenta una alta relación con el problema abordado en el presente trabajo de tesis. El análisis comparativo de los trabajos relacionados se hace en base de los siguientes puntos:

Modelos de regresión lineal: Son aquellos que ayudan a estudiar la relación entre una variable dependiente y una o más variables independientes.

Modelos de predicción: Son aquellos que ayudan a hacer predicciones de una variable.

Evaluación de modelos: Se refiere a la utilización de técnicas para evaluar la eficacia de los modelos generados.

Estudio de contaminantes del aire: Se refiere a que el tema de estudio incluya uno o más contaminantes del aire.

Estudio de problemas de salud: Se refiere a que el tema de estudio incluya uno o más problemas de salud.

En el cuadro 3.1 se desglosan las características presentes que se pueden encontrar en las investigaciones citadas y su relación con la investigación con la que se está trabajando actualmente.

Cuadro 3.1: Comparación de trabajos frente al desarrollado, donde ✓ indica que cumple con esta característica y × no cumple con esta característica.

Trabajo	Modelos de regresión lineal	Modelos de predicción	Evaluación de modelos	Estudio de contaminantes del aire	Estudio de problemas de salud
Martín y Bayle [17]	✓	×	×	✓	✓
Guarnaccia *et al.* [10]	×	✓	✓	✓	×
Julia *et al.* [13]	✓	✓	×	✓	✓
Zhang *et al.* [22]	✓	✓	×	✓	✓
Kim *et al.* [14]	✓	×	×	✓	✓
Liu *et al.* [16]	×	✓	✓	✓	✓
To *et al.* [21]	×	×	×	✓	✓
Breton *et al.* [4]	✓	✓	×	✓	✓
Gupta *et al.* [11]	✓	✓	×	✓	✓
El presente trabajo	✓	✓	✓	✓	✓

3.2.1 Áreas de oportunidad

Como se puede observar en el cuadro 3.1, la mayoría de los trabajos encontrados abordan el estudio de los contaminantes del aire y salud con excepción de Guarnaccia *et al.* [10] que se enfocan en la predicción de niveles de contaminantes del aire, lo cual puede indicar que la relación entre los contaminantes del aire y salud es un tema de relevancia en la actualidad.

Ya que la mayoría de los trabajos encontrados estudian la relación entre contaminantes del aire y salud, la mayoría de los trabajos emplean modelos de regresión lineal por que es una buena opción para el estudio de relaciones entre variables. Las excepciones, además de la ya anteriormente mencionada, son Liu *et al.* [16] y To *et al.* [21] quienes emplean otros tipos de modelos de regresión.

En el presente trabajo se elaboran modelos de predicción para el tratamiento de los datos empleados para los experimentos, ya que como mencionan Zhang *et al.* [22], una de las limitaciones en este tipo de estudios es los campos sin llenar en los registros de datos.

En el presente trabajo también se emplean técnicas para evaluar los modelos generados. Solo en tres de los trabajos encontrados se aborda la evaluación de los modelos empleados, y al ser incluida en el presente estudio, puede representar una distinción.

Capítulo 4

Solución propuesta

En este capítulo se comparte la propuesta de diseño de la solución para el problema de investigación abordado en este presente trabajo, así como su implementación.

4.1 Diseño de la solución propuesta

En diseño de la solución propuesta se plantean las herramientas utilizadas y los pasos seguidos para la solución propuestas.

Las herramientas utilizadas en la presente investigación se muestran en el cuadro 4.1.

Cuadro 4.1: Herramientas utilizadas.

Herramienta	Versión	URL
`Python`	3.8.8	`https://www.python.org/`
`Jupyter Notebook`	6.3.0	`https://jupyter.org/`
`Imageio`	2.9.0	`https://imageio.readthedocs.io/`
`Latextable`	0.2.1	`https://pypi.org/project/latextable/`
`Matplotlib`	3.3.4	`https://matplotlib.org/`
`NumPy`	1.20.1	`http://www.numpy.org/`
`Pandas`	1.2.4	`https://pandas.pydata.org/`
`Seaborn`	0.11.1	`https://seaborn.pydata.org/`
`Scikit-learn`	0.24.1	`https://scikit-learn.org/`
`SciPy`	1.6.2	`https://docs.scipy.org/`
`Statsmodels`	0.12.2	`https://www.statsmodels.org/`
`Texttable`	1.6.4	`https://pypi.org/project/texttable/`

4.1.1 Recolección de datos

La primera fase es la recolección de datos. El objetivo es tener un archivo que contenga datos de los niveles de uno o más contaminantes del aire en años recientes y también del mismo lugar contar con datos del número de egresos hospitalarios durante esos años.

4.1.2 Selección y agrupación de datos

Después de la recolección de datos se seleccionan cuáles van a ser utilizados para los experimentos. Para ello se utiliza `Python` con la librería `Pandas` que permite la manipulación de datos. Para la selección y agrupación de datos en semanas epidemiológicas se sigue el procedimiento mostrado en el algoritmo 1.

Algoritmo 1 Selección y agrupamiento de datos

1: $a \leftarrow$ años de los que se obtuvieron datos
2: **for** $i \in a$ **do**
3: *contaminantes* $\leftarrow$ nombre del archivo .csv que contiene los datos de los contaminantes en el año i
4: Leer en *contaminantes* las columnas fecha y contaminante
5: *egresos* $\leftarrow$ nombre del archivo .csv que contiene los datos de los contaminantes en el año i
6: Leer en *contaminantes* las columnas fecha, padecimiento y estado
7: *estado* $\leftarrow$ estado del que se quieren obtener datos
8: Seleccionar en *contaminantes* los datos del *estado*
9: **end for**

4.1.3 Visualización de la evolución de las variables

Al ya tener seleccionados los datos a utilizar se procede a elaborar gráficos en `Python` que muestran la evolución de las variables en el tiempo. Para ello se generan los tipos de gráficos discutidos a continuación.

4.1.3.1 Series de tiempo

Se realizan series de tiempo en `Python` con ayuda de la librería `Matplotlib`, `Scikit-learn` y `Seaborn`, ya que son herramientas accesibles que ayudan a la generación de este tipo de gráficos.

4.1.3.2 Gráficos de radar

Los gráficos de radar o diagramas de telaraña son otra manera de visualizar un conjunto de datos. Sirven para representar datos en un diagrama bidimensional cuando hay múltiples variables y para comparar variables visualizando si existen valores o patrones de evolución en el tiempo similares entre ellas. Se puede utilizar para graficar un grupo de variables o para comparar varios grupos de las mismas variables. Cada variable se grafica en ejes que se organizan equitativamente alrededor

de un punto de origen central. Las líneas de cuadrícula conectan los ejes y al conectar cada grupo de variables en el gráfico, las conexiones forman un polígono.

Es por ello que en el presente trabajo se elaboran gráficos de radar con ayuda de `Python` y las librerías `NumPy` y `Matplotlib`. En la figura 4.1 se muestra un ejemplo de los gráficos de telaraña generados. Para su interpretación: se identifica la categoría que representa cada eje, se determina cómo se relaciona cada categoría con las demás, se observa la forma completa creada, se lee alrededor de la rueda y se comparan los datos.

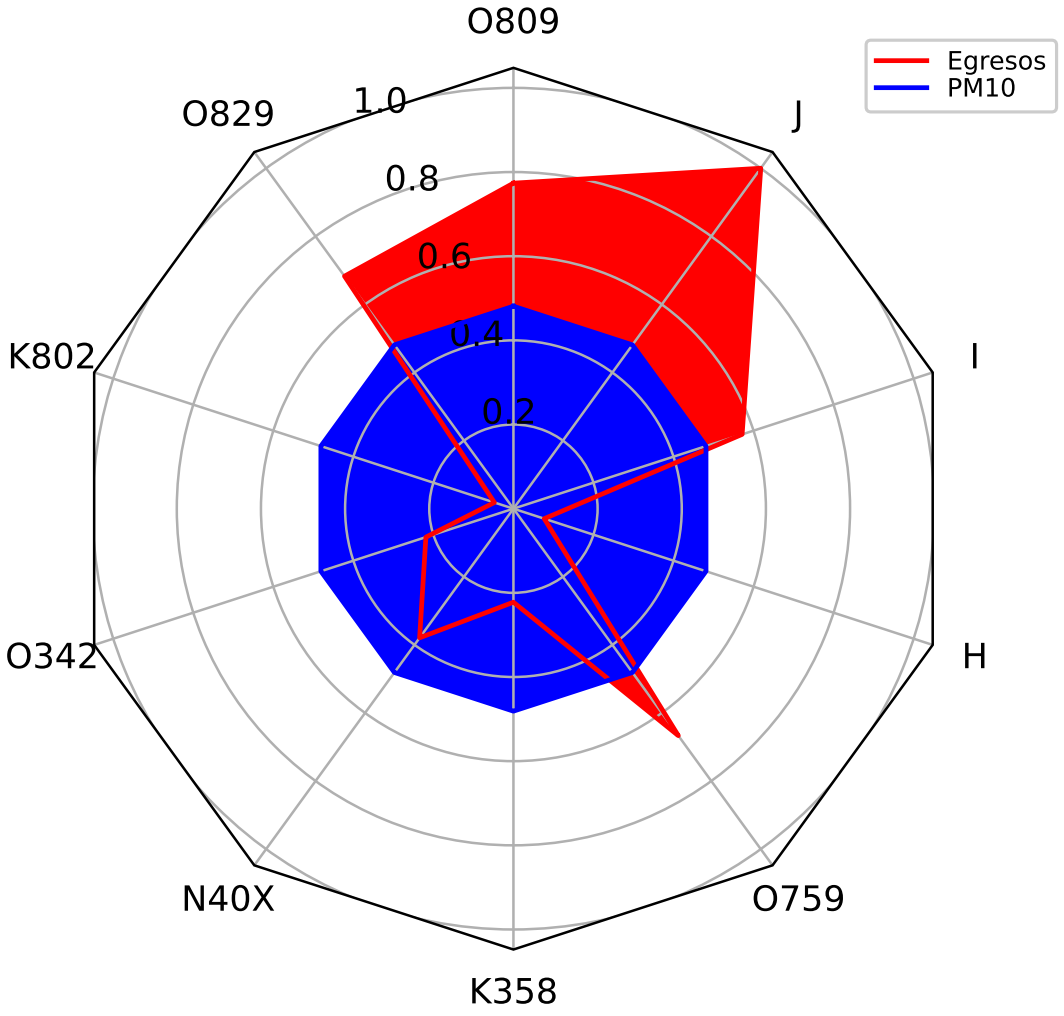


Figura 4.1: Nivel del contaminante PM10 y egresos por CIE en la semana 1 del año 2018.

4.1.4 Implementación de modelos

Después de haber generado gráficos para la visualización de la evolución de las variables, se procede a generar modelos para el estudio de la relación entre las variables. Para ello se utiliza `Python` y la librería `Statsmodels`. Los tipos de modelos generados se discuten a continuación.

4.1.4.1 Regresión lineal

Primeramente se calcula el coeficiente de correlación de Pearson y se verifica que esté entre -1 y 1. Si un valor es cercano a 0, quiere decir que no hay dependencia lineal. Si no hay una dependencia lineal no existe sustento para el modelo de regresión lineal. El modelo de regresión lineal arroja un valor de R^2 que indica en qué grado la variable independiente explica la varianza de la variable dependiente. Además, se obtiene el valor p que indica la relevancia del resultado y se obtiene la raíz de error cuadrático medio (RMSE) que indica cuántas unidades se alejan los valores predichos por el modelo de los valores reales, eso ayuda a determinar el error del modelo.

4.1.4.2 Regresión lineal múltiple

En los modelos de regresión lineal múltiple se tiene más de una variable independiente. Primero se calcula el coeficiente de correlación de Pearson y se verifica que hay una dependencia lineal para generar el modelo de regresión lineal múltiple. El modelo arroja un valor de R^2 que indica en qué grado las variables independientes explican la varianza de la variable dependiente. Además, se obtiene el valor p que indica la relevancia del resultado y se obtiene la raíz de error cuadrático medio (RMSE) que indica cuántas unidades se alejan los valores predichos por el modelo de los valores reales, eso ayuda a determinar el error del modelo.

4.2 Implementación de la solución propuesta

En implementación de la solución propuesta se muestra el desarrollo realizado de los puntos planteados en la sección 4.1. La figura 4.2 muestra las fases seguidas para el desarrollo de la solución propuesta, la cual consiste en la generación de visualizaciones y modelos. El desarrollo del presente proyecto se encuentra en el siguiente repositorio de Github: `https://github.com/selenebpradop/relaciones-contaminantes-salud/`.

En el fragmento de código 4.1 se muestra el proceso realizado para el procesamiento y agrupamiento de los datos en semanas epidemiológicas, esto para que los datos puedan ser utilizados para generar figuras y modelos de regresión lineal.

El fragmento de código 4.2 muestra cómo es que se generan las series de tiempo, esto después de haber procesado y agrupado los datos.

El fragmento de código mostrado en 4.3 genera una animación de gráficos de radar al ingresarle como parámetros los datos ya procesados y agrupados. Dicha animación es generada en formato .mp4 o .gif para poder visualizar la evolución de las variables por semana del año y contaminante.

En el fragmento de código 4.4 se muestra cómo son generados los modelos de regresión lineal después de obtener un coeficiente de correlación de Pearson entre -1 y 1.

El fragmento de código 4.5 muestra cómo se generan los modelos de regresión lineal múltiple después de generar los modelos de regresión lineal individuales. Todos los modelos generados por librería `Statsmodels` se guardan en formato .tex.

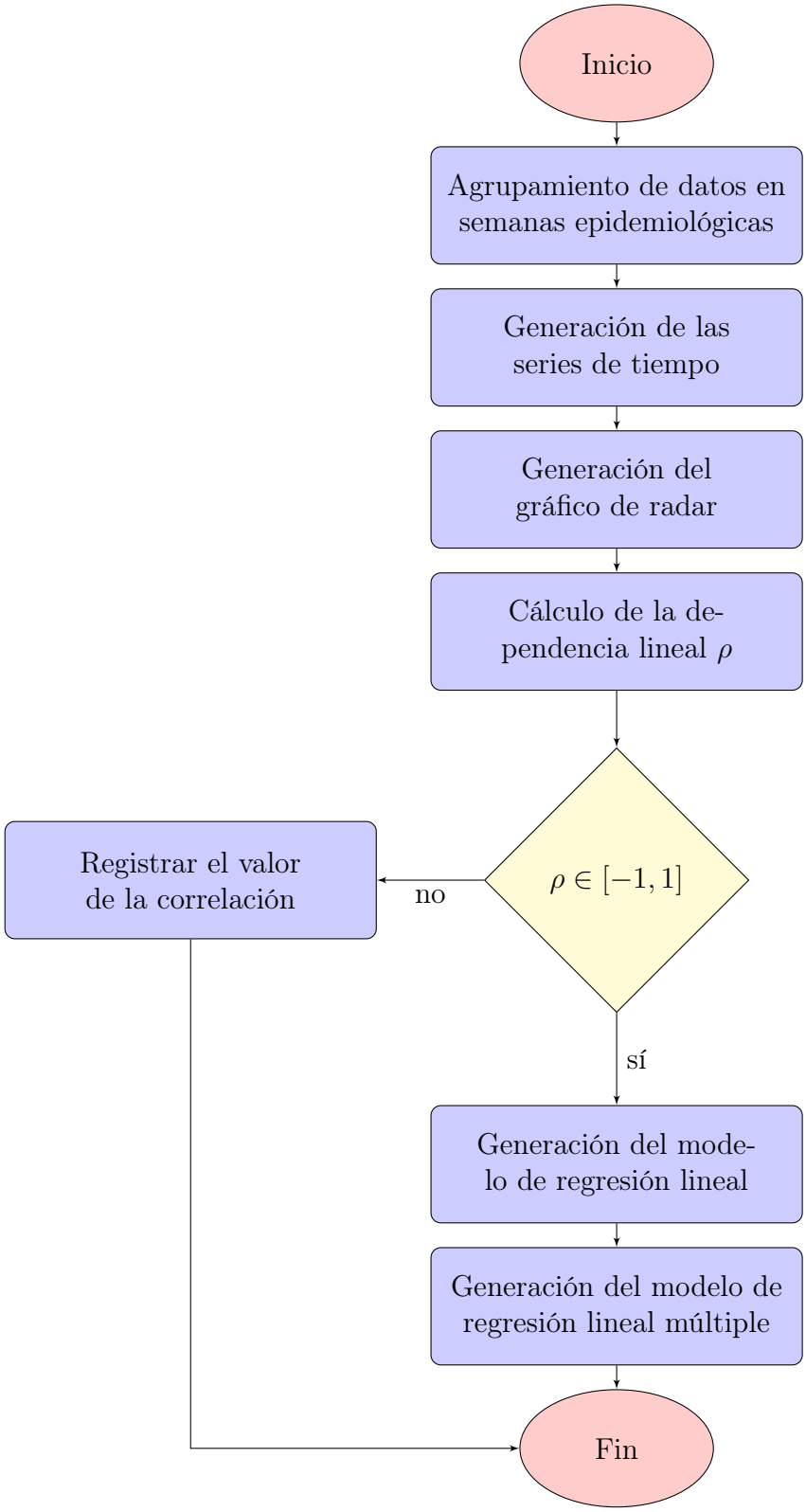


Figura 4.2: Fases del desarrollo de la solución.

```
316 import pandas as pd
317 from epiweeks import Week, date
318 from sklearn import preprocessing
319 import seaborn as sns
320 import matplotlib.pyplot as plt
321 import string
322
323 columns = ['timestamp', contaminante]
324 dataframec = pd.read_csv('filled.csv', usecols=columns).dropna()
325 strfdt = '%d-%b-%y %H'
326 dataframec['timestamp'] = pd.to_datetime(dataframec['timestamp'], errors = 'coerce', format=
327     strfdt)
328 dataframec = dataframec.dropna()
329 dataframec = dataframec.reset_index(drop=True)
330 dataframec['timestamp'] = dataframec['timestamp'].apply(lambda x: x.strftime('%Y-%m-%d %H'))
331 dataframeca = dataframec.loc[dataframec['timestamp'].str.startswith(año)]
332 dataframeca = dataframeca.reset_index(drop=True)
333 strfdt = '%Y-%m-%d %H'
334 dataframeca['timestamp'] = pd.to_datetime(dataframeca['timestamp'], errors = 'coerce', format=
335     strfdt)
336 dataframeca['sem'] = dataframeca['timestamp'].apply(lambda x: date(x.year, x.month, x.day))
337 dataframeca['sem'] = dataframeca['sem'].apply(lambda x: Week.fromdate(x))
338 dataframeca['sem'] = dataframeca['sem'].apply(lambda x: x.week)
339 colums = ['EGRESO', 'DIAG_INI']
340 csvegresos = 'EGRESO_' + año + '.csv'
341 dataframeea = pd.read_csv(csvegresos, usecols=colums).dropna()
342 dataframeea['EGRESO'] = pd.to_datetime(dataframeea['EGRESO'], errors = 'coerce', format=strfdto)
343 dataframeea = dataframeea.loc[dataframeea['ENTIDAD'] == entidad]
344 dataframeea = dataframeea.dropna()
345 dataframeea = dataframeea.reset_index(drop=True)
346 numaño = int(año)
347 dataframeea['sem'] = dataframeea['EGRESO'].apply(lambda x: date(x.year, x.month, x.day))
348 dataframeea['sem'] = dataframeea['sem'].apply(lambda x: Week.fromdate(x))
349 dataframeea['sem'] = dataframeea['sem'].apply(lambda x: x.week)
350 dataframeea['EGRESO'] = dataframeea['EGRESO'].apply(lambda x: x if(x.year==numaño) else pd.NaT)
351 dataframeea = dataframeea.dropna()
352 dataframeea = dataframeea.reset_index(drop=True)
353 dataframesca = pd.DataFrame()
354 dataframesca['sem'] = semanas.index
355 dataframesca[contaminante] = ''
356 n = len(semanas.index)
357 for i in range (n):
358     registrossem = dataframeca.loc[dataframeca['sem'] == i+1]
359     promediocas = registrossem[contaminante].mean()
360     dataframesca[contaminante][i] = promediocas
```

Código 4.1: Procesamiento y agrupamiento de datos.

```
import pandas as pd
from epiweeks import Week, date
from sklearn import preprocessing
import seaborn as sns
import matplotlib.pyplot as plt
import string

diagnosticosaño = dataframeea['DIAG_INI'].value_counts()
diagnosticosaño = diagnosticosaño.sort_values(ascending = False)
ciesaño = dataframeea.groupby(['DIAG_INI', 'sem']).count()

s_scaler = preprocessing.StandardScaler()
ind = []
n = len(semanas.index)
for i in range (n):
    ind.append(i+1)
letras = []
for letra in string.ascii_uppercase:
    letras.append(str(letra))
# Se inicia un contador para controlar la cantidad de graficos a generar
cont = 0
maximo = 10
mindividuales = 7

# Proceso de generación de las figuras
print('\n' + año)
for name in diagnosticosaño.index:
    if cont < maximo:
        dataframegraficoacc = pd.DataFrame()
        dataframegraficoacc[contaminante] = dataframesca[contaminante]
        dataframegraficoacc = dataframegacc.reindex(ind)
        if cont < mindividuales:
            dataframegacc[name] = ciesaño['EGRESO'][name]
            for i in range (n):
                dataframegacc[contaminante][i+1] = dataframesca[contaminante][i]
            col_names = [contaminante, name]
        else:
            nameg =  letras[cont]
            ciesagrupadas = dataframeea.loc[dataframeea['DIAG_INI'].str.startswith(nameg)]
            ciesagrupadas = ciesagrupadas['sem'].value_counts()
            dataframegacc[nameg] = ciesagrupadas
            for i in range (n):
                dataframegacc[contaminante][i+1] = dataframesca[contaminante][i]
            col_names = [contaminante, nameg]
        df_s = s_scaler.fit_transform(dataframegacc)
        df_s = pd.DataFrame(df_s, columns=col_names)
        fig, ax = plt.subplots(ncols=1, figsize=(20, 8))
        print('\n' + col_names[0] + ' & ' + col_names[1])
        ax.set_title('Contaminante ' + col_names[0] + ' & CIE ' + col_names[1])
        ax.set_xlabel('Semana del año ' + año)
        sns.kdeplot(data=df_s)
        plt.savefig(contaminante + '/' + col_names[0] + '&' + col_names[1] + '_' + año + '.jpg',
    format='jpg')
        plt.show()
    cont = cont+1
```

Código 4.2: Generación de series de tiempo.

```
416 def create_spiderwebs(datasets, labels, lenlines, title, titles, spoke_labels, colors, typeframe,
417       outputtype):
418 
419     N1 = len(datasets)
420     N2 = len(labels)
421     N = int(N1/N2)
422     theta = radar_factory(N, frame=typeframe)
423     i=0
424     filenames = []
425     for titlespiderweb in titles:
426         fig, axs = plt.subplots(figsize=(8, 8), subplot_kw=dict(projection='radar'))
427         fig.subplots_adjust(wspace=0.5, hspace=0.20, top=0.85, bottom=0.05)
428         ax = axs
429         ax.set_title(titlespiderweb, weight='bold', size='medium', position=(0.5, 0.5),
430      horizontalalignment='center', verticalalignment='center', fontsize=16)
431         for y in range(N2):
432             dataspider = []
433             xx = y
434             for yy in range(N):
435                 currentdata = datasets[xx]
436                 number = currentdata[i]
437                 nmin = min(currentdata);
438                 nmax = max(currentdata);
439                 r = nmax - nmin
440                 x = (number-nmin)/r
441                 yyy = lenlines*x
442                 dataspider.append(yyy)
443                 xx = xx + N2
444             ax.plot(theta, dataspider, color=colors[y])
445             ax.fill(theta, dataspider, facecolor=colors[y], alpha=0.05)
446         ax.set_varlabels(spoke_labels)
447         ax = axs
448         legend = ax.legend(labels, loc=(0.9, .95), labelspacing=0.1, fontsize='small')
449         i=i+1
450         filename = 'spiderweb' + '_' + title + '_' + str(i) + '.jpg'
451         filenames.append(filename)
452         plt.savefig(filename, format='jpg')
453 
454     # Generate a GIF
455     if(outputtype == 'gif'):
456         with imageio.get_writer(title + '.gif', mode='I', duration=1) as writer:
457             for filename in filenames:
458                 image = imageio.imread(filename)
459                 writer.append_data(image)
460     # Generate a .mp4 video
461     if(outputtype == 'video'):
462         img_array = []
463         for filename in filenames:
464             img = cv2.imread(filename)
465             height, width, layers = img.shape
466             size = (width,height)
467             img_array.append(img)
468         out = cv2.VideoWriter(title + '.mp4',cv2.VideoWriter_fourcc(*'MP4V'), 1, size)
469         for i in range(len(img_array)):
470             out.write(img_array[i])
471         out.release()
```

Código 4.3: Generación de graficos de radar.

```
472 # Gráfico
473 fig, ax = plt.subplots(figsize=(6, 3.84))
474 datos.plot(
475     x    = col_names[0],
476     y    = col_names[1],
477     c    = 'firebrick',
478     kind = "scatter",
479     ax   = ax
480 )
481 ax.set_title('Contaminante ' + col_names[0] + ' & CIE ' + col_names[1])
482 # Correlación lineal entre las dos variables
483 corr_test = pearsonr(x = datos[col_names[0]], y =  datos[col_names[1]])
484 # División de los datos en train y test
485 X = datos[[col_names[0]]]
486 y = datos[col_names[1]]
487 X_train, X_test, y_train, y_test = train_test_split(
488     X.values.reshape(-1,1),
489     y.values.reshape(-1,1),
490     train_size   = 0.8,
491     random_state = 1234,
492     shuffle      = True
493 )
494 X_train = sm.add_constant(X_train, prepend=True)
495 modelo = sm.OLS(endog=y_train, exog=X_train)
496 modelo = modelo.fit()
497 if(modelo.pvalues[1]>0.05):
498     exclude_p.append(col_names[1])
499 # Intervalos de confianza para los coeficientes del modelo
500 modelo.conf_int(alpha=0.05)
501 # Predicciones con intervalo de confianza del 95%
502 pred = modelo.get_prediction(exog = X_train).summary_frame(alpha=0.05)
503 pred.head(4)
504 # Predicciones con intervalo de confianza del 95%
505 pred = modelo.get_prediction(exog = X_train).summary_frame(alpha=0.05)
506 pred['x'] = X_train[:, 1]
507 pred['y'] = y_train
508 pred = pred.sort_values('x')
509 # Gráfico del modelo
510 fig, ax = plt.subplots(figsize=(6, 3.84))
511 ax.set_title('Contaminante ' + col_names[0] + ' & CIE ' + col_names[1])
512 ax.set_xlabel(col_names[0])
513 ax.set_ylabel(col_names[1])
514 ax.scatter(pred['x'], pred['y'], marker='o', color = "gray")
515 ax.plot(pred['x'], pred["mean"], linestyle='-', label="OLS")
516 ax.plot(pred['x'], pred["mean_ci_lower"], linestyle='--', color='red')
517 ax.plot(pred['x'], pred["mean_ci_upper"], linestyle='--', color='red')
518 ax.fill_between(pred['x'], pred["mean_ci_lower"], pred["mean_ci_upper"], 0.1)
519 ax.legend()
520 # Error de test del modelo
521 X_test = sm.add_constant(X_test, prepend=True)
522 pred = modelo.predict(exog = X_test)
523 rmse = mean_squared_error(
524     y_true  = y_test,
525     y_pred  = pred,
526     squared = False
527 )
```

Código 4.4: Generación de los modelos de regresión lineal.

```
528 corr_matrix = datarlm.corr(method='pearson')
529 corr_mat = corr_matrix.stack().reset_index()
530 corr_mat.columns = ['variable_1','variable_2','r']
531 corr_mat = corr_mat.loc[corr_mat['variable_1'] != corr_mat['variable_2'], :]
532 corr_mat['abs_r'] = np.abs(corr_mat['r'])
533 corr_mat = corr_mat.sort_values('abs_r', ascending=False)
534 tidy_corr_matrix = (corr_matrix).head(10)
535 # Heatmap matriz de correlaciones
536 fig, ax = plt.subplots(nrows=1, ncols=1, figsize=(4, 4))
537 sns.heatmap(
538     corr_matrix,
539     annot     = True,
540     cbar      = False,
541     annot_kws = {"size": 8},
542     vmin      = -1,
543     vmax      = 1,
544     center    = 0,
545     cmap      = sns.diverging_palette(20, 220, n=200),
546     square    = True,
547     ax        = ax)
548 ax.set_xticklabels(
549     ax.get_xticklabels(),
550     rotation = 45,
551     horizontalalignment = 'right',)
552 ax.tick_params(labelsize = 10)
553 # División de los datos en train y test
554 X = datarlm[spoke_labels]
555 y = datarlm[contaminante]
556 X_train, X_test, y_train, y_test = train_test_split(
557     X,
558     y.values.reshape(-1,1),
559     train_size   = 0.8,
560     random_state = 1234,
561     shuffle      = True)
562 # Creación del modelo utilizando matrices como en scikitlearn
563 X_train = sm.add_constant(X_train, prepend=True)
564 modelo = sm.OLS(endog=y_train, exog=X_train,)
565 modelo = modelo.fit()
566 sml = modelo.summary().as_latex()
567 namefile = 'modelos_latex/' + 'regresion_lineal_multiple_' + contaminante + '_' + año + '.tex'
568 f = open(namefile, 'w')
569 with open(namefile, 'w') as f:
570     f.write(sml)
571 # Diagnóstico errores (residuos) de las predicciones de entrenamiento
572 y_train = y_train.flatten()
573 prediccion_train = modelo.predict(exog = X_train)
574 residuos_train   = prediccion_train - y_train
575 # Predicciones con intervalo de confianza
576 predicciones = modelo.get_prediction(exog = X_train).summary_frame(alpha=0.05)
577 predicciones.head(4)
578 # Error de test del modelo
579 X_test = sm.add_constant(X_test, prepend=True)
580 rmse = mean_squared_error(
581     y_true  = y_test,
582     y_pred  = modelo.predict(exog = X_test),
583     squared = False)
```

Código 4.5: Generación de los modelos de regresión lineal múltiple.

Capítulo 5

Experimentos

En el presente capítulo se presenta el diseño de los experimentos realizados así como los resultados obtenidos de ellos.

En esta sección se tratan los resultados obtenidos partiendo de desarrollar algunos experimentos que permiten determinar si la solución propuesta cumple con el objetivo planteado.

5.1 Diseño experimental

En la presente sección se discute el diseño de experimentos, es decir, qué valores constantes fueron utilizados para su realización y por qué se usan dichos valores.

5.1.1 Datos de entrada

Los datos de egresos hospitalarios de los años 2017 y 2018 provienen de la base de datos de la Secretaría de Salud del Gobierno de México [9]. También se tienen registros de los niveles de PM10 y PM2.5 presentes en el área metropolitana de Monterrey, dichos registros son obtenidos por las estaciones de monitoreo pertenecientes

al SIMA [20] mostradas en la figura 1.1 en la página 2. Los documentos con los datos son proporcionados por Benavides [3].

Selección de datos. Los conjuntos de datos por año de egresos hospitalarios contienen información de todos los estados de México, por lo cual se hace una limpieza de datos para solo obtener los registros de Nuevo León ya que de dicha entidad es de la cual se tienen los datos de contaminación.

Datos de ingresos hospitalarios. Se agrupan en semanas epidemiológicas para una mejor manipulación de ellos. Por CIE se obtiene el número de egresos en cada semana del año ordenadas por el número de egresos en el año partiendo de la CIE con la mayor cantidad.

Datos de los contaminantes. Se agrupan en semanas epidemiológicas para una mejor manipulación de ellos. Se obtiene el promedio del nivel del contaminante por cada semana del año.

5.1.2 Visualización de datos

Con los datos ya seleccionados y agrupados se procede a generar las visualizaciones de los datos. Las visualizaciones de los datos se hacen con series de tiempo y gráficos de radar.

Series de tiempo. Se procede a generar series de tiempo de cada año por contaminante y CIE. Las variables ajustables son:

- El nombre del contaminante.
- El año del que se quieren obtener las series de tiempo.
- Número de series de tiempo a generar por contaminante. Se parte de la CIE con mayor número de egresos.

Gráficos de radar. Los datos ya seleccionados y agrupados se normalizan teniendo como valor mínimo cero y como valor máximo un número entre uno y cuatro. Posteriormente se generan gráficos de radar de cada año por semana, en las que se muestra el nivel contaminante y la variación de las CIE. Las variables ajustables son:

- Cantidad de CIE a agregar en el gráfico. Se parte de la CIE con mayor número de egresos.
- Valor máximo que se utiliza para representar la longitud de los ejes en el gráfico.
- Nombre de la figura.
- Nombre de cada eje en el gráfico.
- Los colores de cada eje en el gráfico.
- Si el gráfico es generado de forma circular o en forma de polígono.

5.1.3 Generación de modelos

Se procede a generar los modelos. En cada modelo se tienen métricas para evaluar su eficacia y valores que pueden ser ajustados en función de encontrar la combinación que proporcione mejores resultados.

Se tienen algunas variables que pueden ser modificadas para la generación de los modelos de regresión lineal.

- Cantidad de CIE a agregar en el modelo. Se parte de la CIE con mayor número de egresos.
- Porcentaje de datos utilizados para el entrenamiento del modelo.
- Nivel de significancia.

También se tienen variables que indican información sobre la eficacia del modelo.

- Valor p.
- R^2 (R cuadrado).
- Raíz de error cuadrático medio (RMSE).

5.2 Resultados

Establecidas las especificaciones de los experimentos que se realizan, se reportan los resultados obtenidos. Los experimentos se elaboran por contaminante, desglosando los resultados por año. En la carpeta `https://github.com/selenebpradop/relaciones-contaminantes-salud/tree/main/figuras/` se encuentran animaciones en video de los gráficos de radar generados por contaminante y año y todas las imágenes de las series de tiempo obtenidas.

5.2.1 Experimento A: Datos de niveles de PM10

Se estudian los niveles del contaminante PM10 de los años 2017 y 2018.

5.2.1.1 Año 2017

En la figura 5.1 se muestra una de las series de tiempo generadas para la CIE con mayor número de egresos registrados en el conjunto de datos del año. Además, en el cuadro 5.1 se presentan los resultados obtenidos de los modelos de regresión lineal y la eficacia obtenida de dichos modelos. El cuadro 5.2 muestra los resultados del modelo de regresión lineal múltiple. En la figura 5.2 se muestran las correlaciones obtenidas en un gráfico de radar.

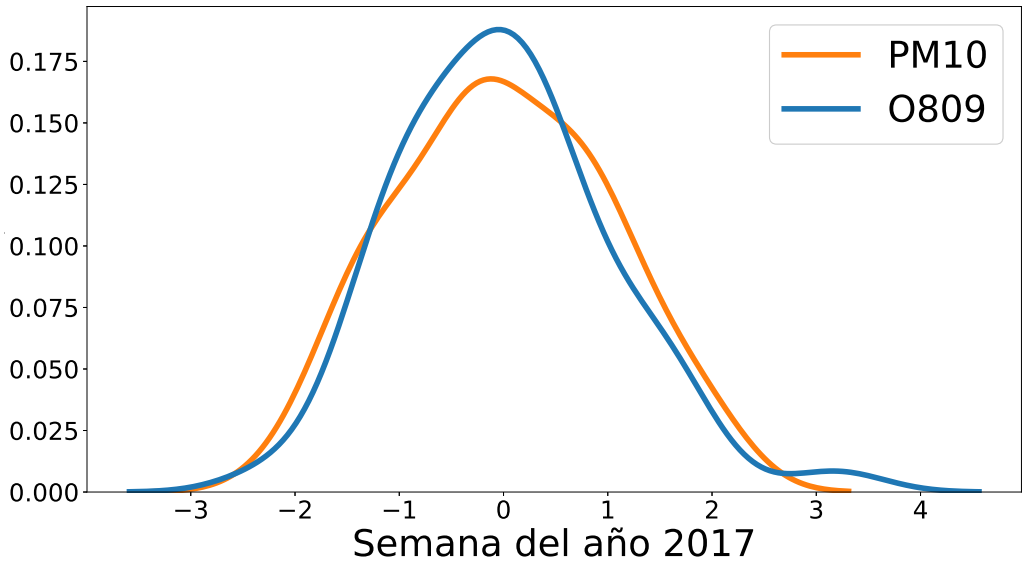


Figura 5.1: Evolución de los niveles de PM10 y el número de egresos diagnosticados con la CIE O809 en el 2017.

Cuadro 5.1: Resultados obtenidos PM10 2017

CIE	ρ	R^2	Valor p	ϵ
O809	-0.275	0.061	0.121	0.239
O829	0.100	0.091	0.055	0.287
O759	-0.085	0.116	0.029	0.294
O069	-0.247	0.070	0.094	0.222
K802	0.044	0.005	0.658	0.282

ρ = Coeficiente de correlación de Pearson
ϵ = RMSE para medir el error

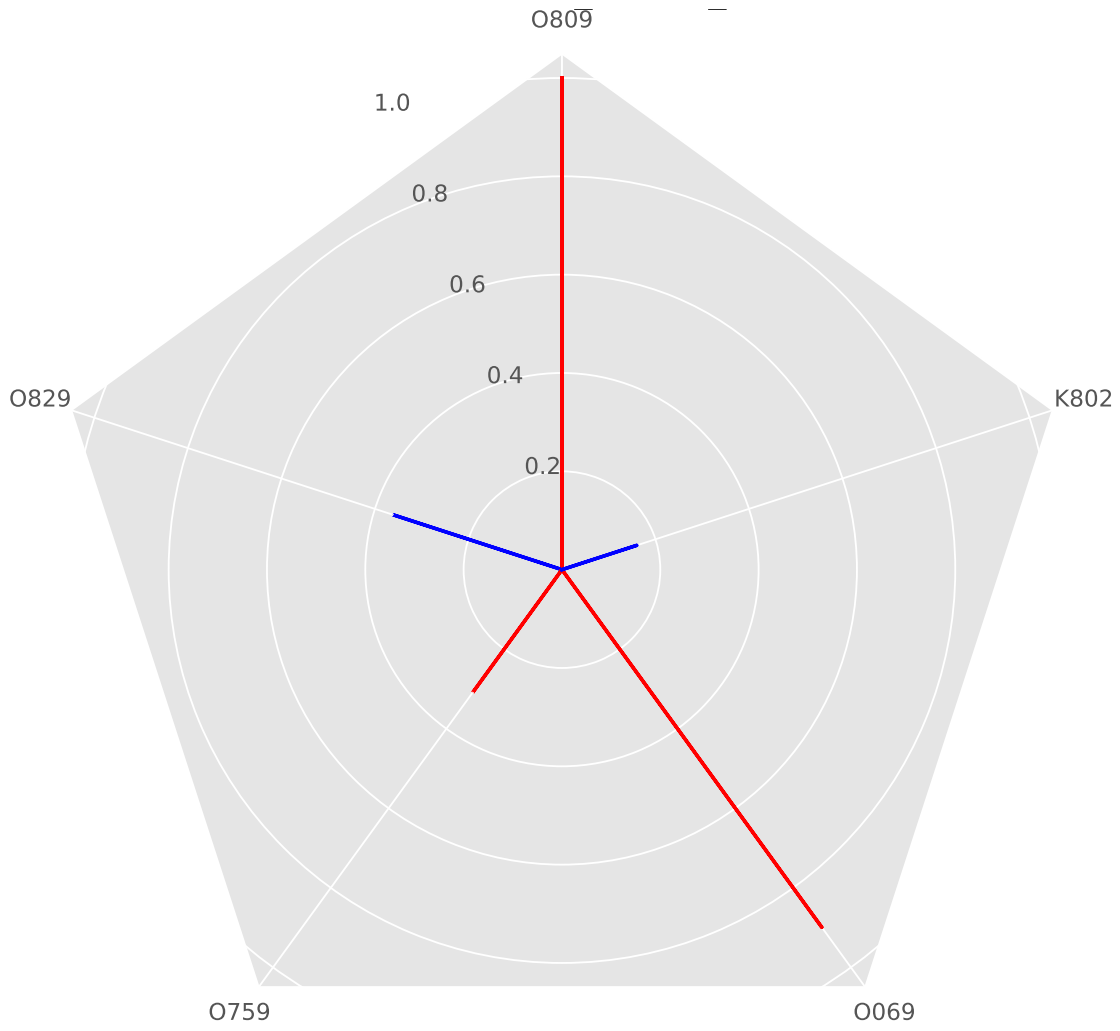


Figura 5.2: Correlaciones entre los niveles de PM10 y CIE en el 2017 donde el azul indica una correlación positiva y el rojo una correlación negativa.

Cuadro 5.2: Resultados regresión lineal múltiple PM10 2017

Variable Dep.:		y	**R^2:**		0.313	
Modelo:		OLS	**Método:**		Mínimos cuadrados	
Error:		0.226				
	coef	**std err**	t	**P> |t|**	**[0.025**	**0.975]**
const	0.5859	0.146	4.002	0.000	0.289	0.883
O809	-0.3019	0.122	-2.472	0.018	-0.550	-0.054
O829	0.2685	0.123	2.185	0.036	0.019	0.518
O759	-0.2229	0.182	-1.222	0.230	-0.593	0.147
O069	-0.1441	0.120	-1.200	0.238	-0.388	0.100
K802	-0.0456	0.133	-0.344	0.733	-0.315	0.224

5.2.1.2 Año 2018

En la figura 5.3 se muestra una de las series de tiempo generadas para la CIE con mayor número de egresos registrados en el conjunto de datos del año. Además, en el cuadro 5.3 se presentan los resultados obtenidos de los modelos de regresión lineal y la eficacia obtenida de dichos modelos. El cuadro 5.4 muestra los resultados del modelo de regresión lineal múltiple. En la figura 5.4 se muestran las correlaciones obtenidas en un gráfico de radar.

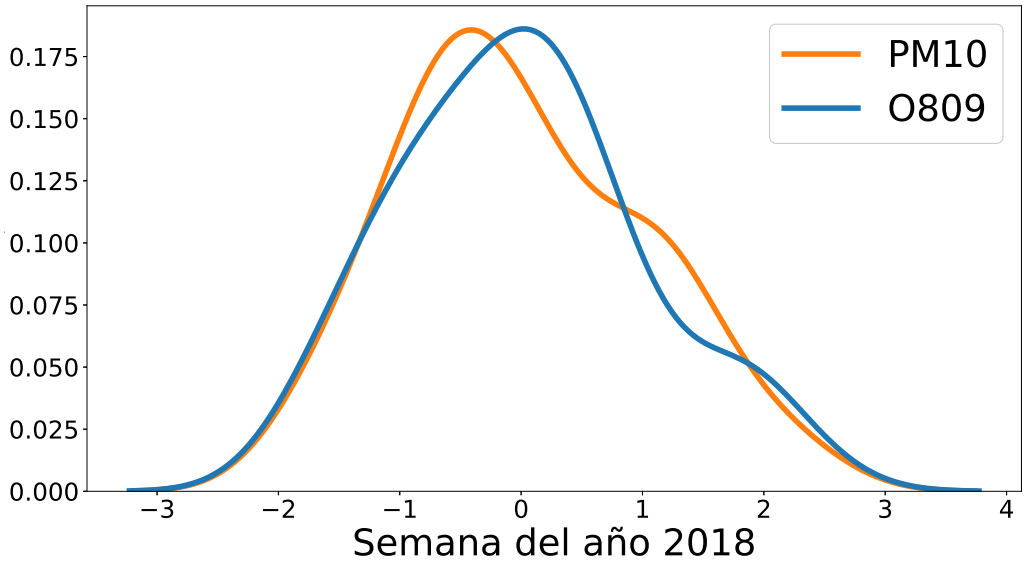


Figura 5.3: Evolución de los niveles de PM10 y el número de egresos diagnosticados con la CIE O809 en el 2018.

Cuadro 5.3: Resultados obtenidos PM10 2018

CIE	ρ	R^2	Valor p	ϵ
O809	-0.271	0.015	0.440	0.275
O829	0.282	0.069	0.098	0.249
K802	0.277	0.084	0.066	0.152
O342	-0.401	0.100	0.044	0.248
N40X	-0.009	0.000	0.964	0.243

ρ = Coeficiente de correlación de Pearson
ϵ = RMSE para medir el error

Cuadro 5.4: Resultados regresión lineal múltiple PM10 2018

Variable Dep.:	y	**R^2:**	0.182	
Modelo:	OLS	**Método:**	Mínimos cuadrados	
Error:	0.159			

	coef	**std err**	t	**P> |t|**	**[0.025**	**0.975]**
const	0.3885	0.210	1.849	0.073	-0.038	0.815
O809	-0.0114	0.188	-0.061	0.952	-0.392	0.370
O829	0.0854	0.214	0.400	0.692	-0.349	0.520
K802	0.3088	0.167	1.852	0.072	-0.030	0.647
O342	-0.1708	0.208	-0.820	0.418	-0.593	0.252
N40X	-0.1239	0.167	-0.740	0.464	-0.464	0.216

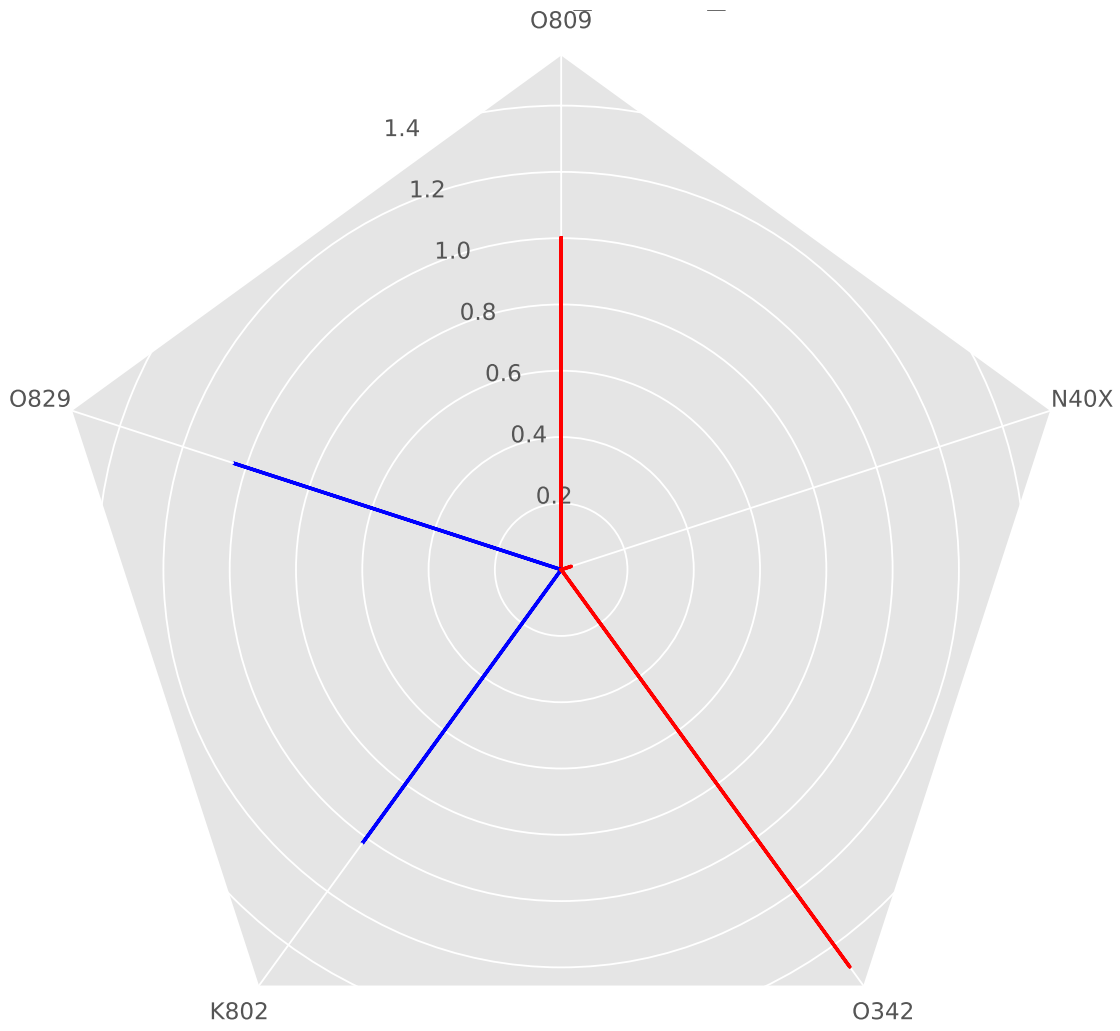


Figura 5.4: Correlaciones entre los niveles de PM10 y CIE en el 2018 donde el azul indica una correlación positiva y el rojo una correlación negativa.

5.2.2 Experimento B: Niveles de PM2.5

Se estudian los niveles del contaminante PM2.5 de los años 2017 y 2018.

5.2.2.1 Año 2017

En la figura 5.5 se muestra una de las series de tiempo generadas para la CIE con mayor número de egresos registrados en el conjunto de datos del año. Además, en el cuadro 5.5 se presentan los resultados obtenidos de los modelos de regresión lineal y la eficacia obtenida de dichos modelos. El cuadro 5.6 muestra los resultados del modelo de regresión lineal múltiple. En la figura 5.6 se muestran las correlaciones obtenidas en un gráfico de radar.

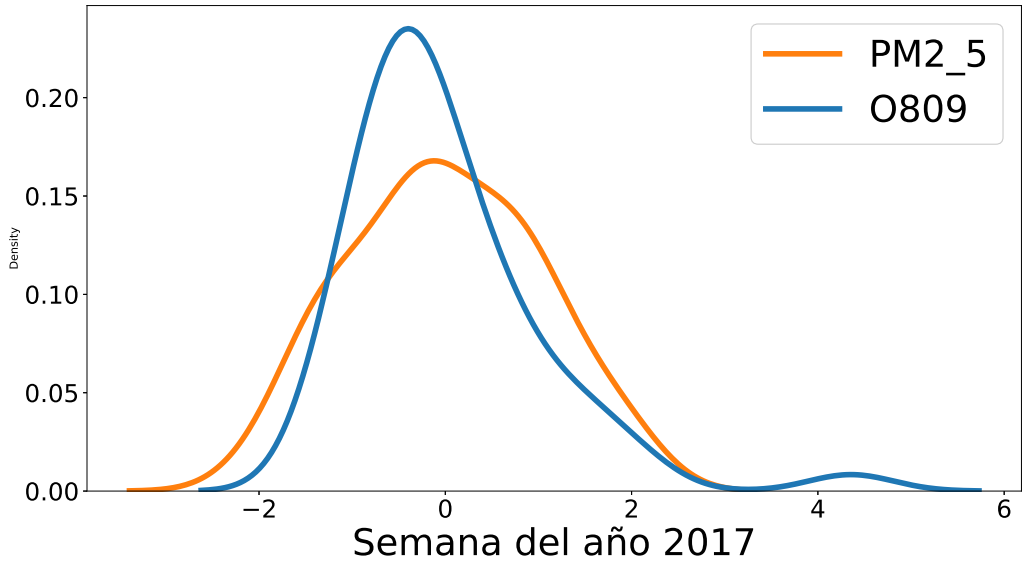


Figura 5.5: Evolución de los niveles de PM2.5 y el número de egresos diagnosticados con la CIE O809 en el 2017.

Cuadro 5.5: Resultados obtenidos PM2.5 2017

CIE	ρ	R^2	Valor p	ϵ
O809	-0.093	0.011	0.511	0.256
O829	0.014	0.021	0.371	0.253
O759	-0.172	0.113	0.032	0.278
O069	-0.350	0.112	0.033	0.208
K802	0.006	0.005	0.667	0.285

ρ = Coeficiente de correlación de Pearson
ϵ = RMSE para medir el error

Cuadro 5.6: Resultados regresión lineal múltiple PM2.5 2017

Variable Dep.:		y	**R^2:**		0.210	
Modelo:		OLS	**Método:**		Mínimos cuadrados	
Error:		0.175				
	coef	**std err**	t	**P> |t|**	**[0.025**	**0.975]**
const	0.5345	0.158	3.373	0.002	0.213	0.856
O809	-0.1754	0.132	-1.327	0.193	-0.444	0.093
O829	0.0701	0.133	0.527	0.602	-0.200	0.340
O759	-0.2585	0.197	-1.309	0.199	-0.659	0.142
O069	-0.2148	0.130	-1.652	0.108	-0.479	0.049
K802	-0.1164	0.144	-0.811	0.423	-0.408	0.175

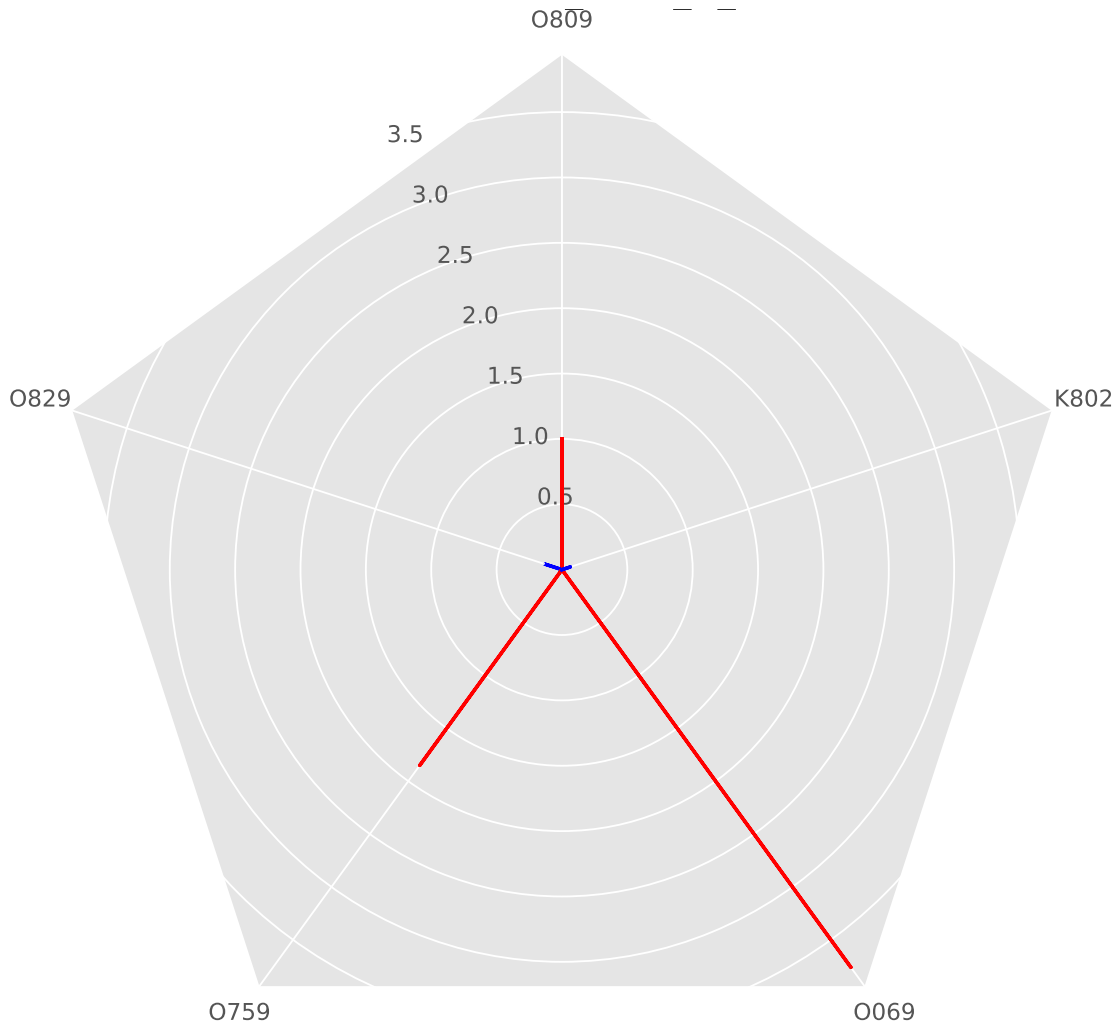


Figura 5.6: Correlaciones entre los niveles de PM2.5 y CIE en el 2017 donde el azul indica una correlación positiva y el rojo una correlación negativa.

5.2.2.2 Año 2018

En la figura 5.7 se muestra una de las series de tiempo generadas para la CIE con mayor número de egresos registrados en el conjunto de datos del año. Además, en el cuadro 5.7 se presentan los resultados obtenidos de los modelos de regresión lineal y la eficacia obtenida de dichos modelos. El cuadro 5.8 muestra los resultados del modelo de regresión lineal múltiple. En la figura 5.8 se muestran las correlaciones obtenidas en un gráfico de radar.

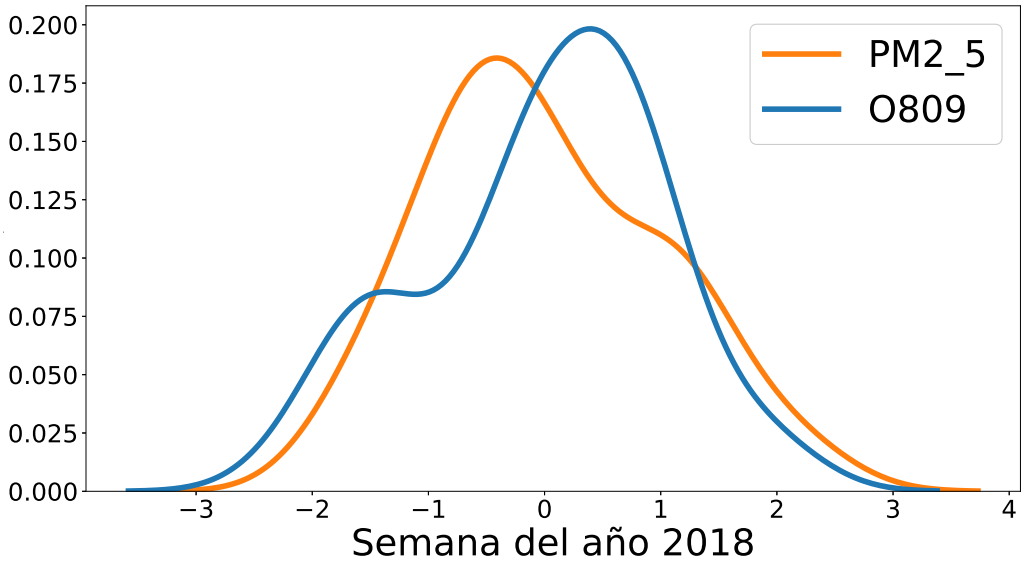


Figura 5.7: Evolución de los niveles de PM2.5 y el número de egresos diagnosticados con la CIE O809 en el 2018.

Cuadro 5.7: Resultados obtenidos PM2.5 2018

CIE	ρ	R^2	Valor p	ϵ
O809	-0.354	0.057	0.133	0.259
O829	0.225	0.046	0.177	0.256
K802	0.273	0.089	0.058	0.159
O342	-0.443	0.149	0.013	0.245
N40X	-0.074	0.001	0.842	0.241

ρ = Coeficiente de correlación de Pearson
ϵ = RMSE para medir el error

Cuadro 5.8: Resultados regresión lineal múltiple PM2.5 2018

Variable Dep.:	y	**R^2:**	0.259
Modelo:	OLS	**Método:**	Mínimos cuadrados
Error:	0.174		

	coef	**std err**	t	**P> |t|**	**[0.025**	**0.975]**
const	0.7042	0.193	3.652	0.001	0.313	1.096
O809	-0.0863	0.172	-0.501	0.620	-0.436	0.264
O829	-0.1084	0.196	-0.553	0.584	-0.507	0.290
K802	0.3006	0.153	1.964	0.058	-0.010	0.611
O342	-0.3311	0.191	-1.733	0.092	-0.719	0.057
N40X	-0.2053	0.154	-1.336	0.190	-0.517	0.107

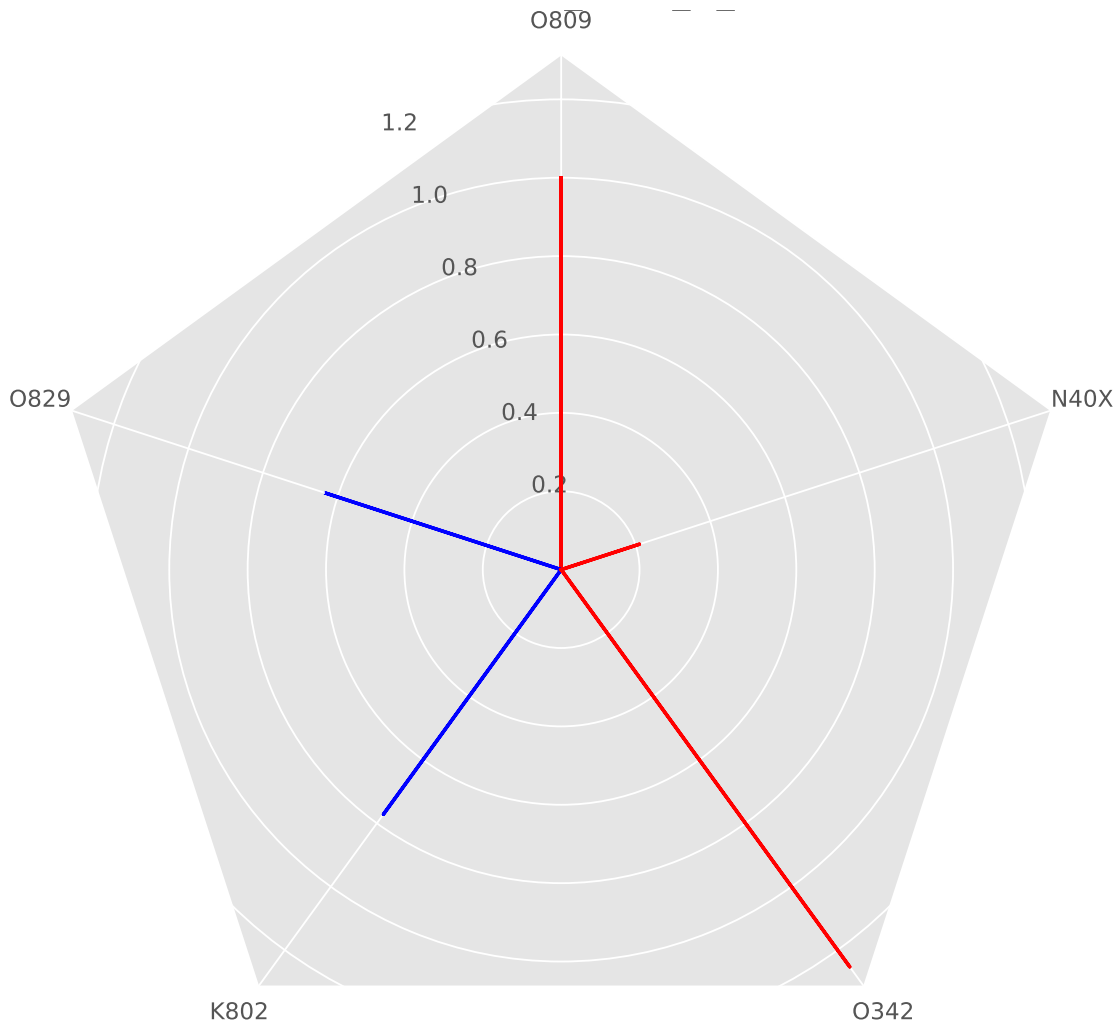


Figura 5.8: Correlaciones entre los niveles de PM2.5 y CIE en el 2018 donde el azul indica una correlación positiva y el rojo una correlación negativa.

5.3 Discusión

Como se puede observar, en todos los experimentos se obtiene una correlación entre -1 y 1 y diferente a 0, por lo tanto se pueden generar los modelos de regresión lineal.

En el Experimento A el error RMSE en los modelos de regresión lineal varía entre 0.152 y 0.294, lo cual indica que no todos los resultados alcanzan una fiabilidad mayor al 80 %, excepto en la CIE K802 en el año 2018 en la cual se encuentra el porcentaje de error más bajo. El valor de R^2 más alto en el año 2017 se encuentra en la CIE O759 con un valor p de 0.029, sin embargo, en el modelo de regresión lineal múltiple se encuentra el valor de R^2 más alto para el contaminante PM10 en el año 2017. En el año 2018 el valor de R^2 más alto se encuentra en la CIE O342 con un valor p de 0.044, sin embargo, en el modelo de regresión lineal múltiple se encuentra el valor de R^2 más alto para el contaminante PM10 en el año 2018.

En el Experimento B el error RMSE en los modelos de regresión lineal varía entre 0.159 y 0.278, lo cual indica que no todos los resultados alcanzan una fiabilidad mayor al 80 %, excepto en la CIE K802 en el año 2018 en la cual se encuentra el porcentaje de error más bajo. El valor de R^2 más alto en el año 2017 se encuentra en la CIE O759 con un valor p de 0.032, sin embargo, en el modelo de regresión lineal múltiple se encuentra el valor de R^2 más alto para el contaminante PM2.5 en el año 2017. En el año 2018 el valor de R^2 más alto se encuentra en la CIE O342 con un valor p de 0.013, sin embargo, en el modelo de regresión lineal múltiple se encuentra el valor de R^2 más alto para el contaminante PM2.5 en el año 2018.

711 Todos los experimentos son ejecutados en `Jupyter Notebook` en una laptop
712 con las especificaciones del cuadro 5.9.

Cuadro 5.9: Especificaciones técnicas del equipo de cómputo

Sistema Operativo	macOS Big Sur
Procesador	Apple M1
RAM	8 GB RAM

Capítulo 6

Conclusiones

El presente capítulo describe la tesis a partir de la manera que cumple los objetivos generales y específicos para determinar si la hipótesis se comprueba, trata también del porqué se realizó la tesis.

En el presente proyecto se generaron visualizaciones para el estudio de las relaciones entre determinados contaminantes y determinadas CIE. Además, se generaron modelos de regresión lineal y modelos de regresión lineal múltiple con diferentes contaminantes y diferentes CIE para poder realizar una comparación entre los modelos generados.

En los experimentos se encontró un coeficiente de correlación de Pearson diferente a 0 entre -1 y 1, por lo tanto se pudieron generar los modelos de regresión lineal. Se encontró que al estudiar las CIE de manera agrupada en un modelo de regresión múltiple se obtiene una mejor explicación de la varianza de los niveles de los contaminantes PM10 y PM2.5 frente al modelo de regresión lineal simple. Cinco de los veinticuatro valores de error RMSE reportados son menores a 0.20, lo cual indica que la mayoría de los valores predichos por los modelos se alejaron más de 0.20 unidades de los valores reales, por lo tanto dichos valores de error son mayores a los deseados.

6.1 Contribuciones

Primeramente se encontró que la CIE que reporta mayor número de egresos en Nuevo León, México en los años 2017 y 2018 es la CIE O809. En las series de tiempo se observa que la cantidad de egresos de la mayoría de las CIE estudiadas presentan una línea de evolución similar al contaminante PM10.

Además, se encontró una correlación lineal entre los contaminantes estudiados y las CIE estudiadas, por lo cual el presente proyecto motiva a seguir realizando labores en torno a la investigación de la dependencia lineal que se presenta entre los contaminantes y las CIE.

Finalmente, se observó que para el estudio de las relaciones entre los contaminantes y las CIE se obtuvieron mejores resultados con los modelos de regresión lineal múltiple frente a los modelos de regresión lineal simple, lo cual indica que se puede obtener información relevante si se emplean modelos de regresión lineal múltiple para realizar investigaciones que estudien las relaciones entre los niveles de determinados contaminantes y el número de egresos por determinadas CIE.

6.2 Trabajo a futuro

El presente trabajo brinda algunos aspectos a considerar para realizar un trabajo a futuro, los cuales son: la recolección de más datos de los niveles de contaminantes y de egresos para el estudio de años más recientes, realizar un estudio de las relaciones de los niveles de determinados contaminantes y la cantidad de egresos por CIE empleando modelos de regresión lineal múltiple, la creación de un mapa interactivo donde se pueda observar por región los niveles de los contaminantes y la cantidad de egresos, y la generación de una página web que funcione en conjunto con el desarrollo elaborado para que al ingresar el archivo con los datos se generen las visualizaciones y modelos.

Las oportunidades de mejora detectada para el presente proyecto son: comparar los resultados obtenidos de los modelos regresión lineal múltiple con el mismo modelo pero incrementando el número de variables independientes (CIE), elaborar modelos de regresión no lineal para comparar sus resultados con los resultados de los modelos generados, y utilizar un conjunto de datos más consistente ya que eso favorece a que se tengan resultados más fiables y precisos.

Apéndice A

CIE y sus nombres de enfermedades

Cuadro A.1: CIE mencionadas en los Experimentos y el nombre de la enfermedad.

CIE	Enfermedad
N40-N53	Enfermedades de los órganos genitales masculino
K00-K93	Enfermedades del aparato digestivo
O00-O99	Embarazo, parto y puerperio

Bibliografía

[1] Arias, J. R. (2006), «What is an epidemiological week and why do we use them», *Skeeter*, **66**(1), pág. 7.

[2] Ballester Díez, F., J. M. Tenías y S. Pérez-Hoyos (1999), «Efectos de la contaminación atmosférica sobre la salud: una introducción», *Revista Española de Salud Pública*, **73**(2), págs. 109–121.

[3] Benavides, A. (2022), «Perfil de Github», `https://github.com/jbenavidesv87`.

[4] Breton, R. M. C., J. C. Breton, M. de la Luz Espinosa Fuentes, J. Kahl, A. A. E. Guzman, R. G. Martínez, C. Guarnaccia, R. del Carmen Lara Severino, E. R. Lara y A. B. Francavilla (2021), «Short-Term Associations between Morbidity and Air Pollution in Metropolitan Area of Monterrey, Mexico», *Atmosphere*, **12**(10), pág. 1352. doi: 10.3390/atmos12101352.

[5] Brockwell, P. J., P. J. Brockwell, R. A. Davis y R. A. Davis (2002), *Introduction to time series and forecasting*, Springer. ISBN 9780387216577.

[6] Catalá, F. y E. De Manuel (1998), «Informe SESPAS 1998: La salud pública y el futuro del estado del bienestar», *Granada: EASP*.

[7] Corbitt, R. y R. Corbitt (1990), *Standard Handbook of Environmental Engineering*, McGraw-Hill. ISBN 9780070131583.

[8] DARLINGTON, R. B. y A. F. HAYES (2016), *Regression analysis and linear models: Concepts, applications, and implementation*, Guilford Publications. ISBN 9781462521135.

[9] DIRECCIÓN GENERAL DE INFORMACIÓN EN SALUD (2021), «Egresos Hospitalarios», `http://www.dgis.salud.gob.mx/contenidos/basesdedatos/da_egresoshosp_gobmx.html`.

[10] GUARNACCIA, C., J. G. C. BRETON, R. M. C. BRETON, C. TEPEDINO, J. QUARTIERI y N. E. MASTORAKIS (2018), «ARIMA models application to air pollution data in Monterrey, Mexico», en *AIP Conference Proceedings*, tomo 1, AIP Publishing LLC, págs. 1–5.

[11] GUPTA, A., H. BHERWANI, S. GAUTAM, S. ANJUM, K. MUSUGU, N. KUMAR, A. ANSHUL y R. KUMAR (2021), «Air pollution aggravating COVID-19 lethality? Exploration in Asian cities using statistical models», *Environment, Development and Sustainability*, **23**(4), págs. 6408–6417. doi: 10.1007/s10668-020-00878-9.

[12] HADDAD, R. (1974), «Contaminación del aire. Situación actual en la América Latina y el Caribe», *Informe técnico.*

[13] JULIA, A., D. A. FC LICHTENFELS, K. VAN DER PLAAT, C. C. DE JONG, D. S. VAN DIEMEN, I. POSTMA, C. M. NEDELJKOVIC, N. VAN DUIJN, S. AMIN, M. LA BASTIDE-VAN GEMERT, D. VRIES *et al.* (2018), «Long-term air pollution exposure, genome-wide DNA methylation and lung function in the LifeLines cohort study», *Environmental health perspectives*, **126**(2), pág. 027 004. doi: 10.1289/EHP2045.

[14] KIM, J. S., Z. CHEN, T. L. ALDERETE, C. TOLEDO-CORRAL, F. LURMANN, K. BERHANE y F. D. GILLILAND (2019), «Associations of air pollution, obesity and cardiometabolic health in young adults: The Meta-AIR study», *Environment international*, **133**, págs. 105–180. doi: 10.1016/j.envint.2019.105180.

[15] KORC, M. y R. SÁENZ (1999), «Monitoreo de la calidad del aire en América Latina», *Korc Marcelo E*, págs. 1–22.

[16] LIU, Y., J. PAN, H. ZHANG, C. SHI, G. LI, Z. PENG, J. MA, Y. ZHOU y L. ZHANG (2019), «Short-term exposure to ambient air pollution and asthma mortality», *American journal of respiratory and critical care medicine*, **200**(1), págs. 24–32. doi: 10.1164/rccm.201810-1823OC.

[17] MARTÍN, R. M. y M. S. BAYLE (2018), «Impacto de la contaminación ambiental en las consultas pediátricas de Atención Primaria: estudio ecológico», en *Anales de Pediatría*, tomo 2, Elsevier, págs. 80–85.

[18] ORGANIZATION, W. H. *et al.* (2016), «International Classification of Diseases. 2016», *World Health Organization*.

[19] PENICHE-CAMPS, S. y M. CORTEZ-HUERTA (2020), «La costumbre al envenenamiento: El caso de los contaminantes atmosféricos de la ciudad de Guadalajara, México», *Revista de Ciencias Ambientales*, **54**(2), págs. 1–19. doi: 10.15359/rca.54-2.1.

[20] SIMA (2015), «Sistema Integral de Monitoreo Ambiental», `http://aire.nl.gob.mx`.

[21] TO, T., J. ZHU, D. STIEB, N. GRAY, I. FONG, L. PINAULT, M. JERRETT, A. ROBICHAUD, R. MÉNARD, A. VAN DONKELAAR *et al.* (2020), «Early life exposure to air pollution and incidence of childhood asthma, allergic rhinitis and eczema», *European Respiratory Journal*, **55**(2). doi: 10.1183/13993003.00913-2019.

[22] ZHANG, Z., B. DONG, S. LI, G. CHEN, Z. YANG, Y. DONG, Z. WANG, J. MA y Y. GUO (2019), «Exposure to ambient particulate matter air pollution, blood pressure and hypertension in children and adolescents: a national cross-sectional study in China», *Environment international*, **128**, págs. 103–108. doi: 10.1016/j.envint.2019.04.036.

Resumen autobiográfico

Selene Berenice Prado Prado

Candidato para obtener el grado de
Ingeniería en Tecnología de Software

Universidad Autónoma de Nuevo León
Facultad de Ingeniería Mecánica y Eléctrica

Tesis:
Modelado y visualización de relaciones entre contaminantes del aire y salud pública

Nací el 30 de Junio de 2000 en Monterrey, Nuevo León, soy la mayor de cuatro hijos. Mi familia está conformada por mi madre Lilia Prado López, mi padre Adan Alfaro Lerma, y mis hermanos: Angel Alejandro Prado Prado, Estrella Belen Prado Prado, y Genesis Adali Alfaro Prado.

Desde pequeña me han gustado las matemáticas, aprender cómo funcionan los sistemas computacionales, y leer.

Durante los primeros semestres de mi carrera descubrí la inteligencia computacional, un área que me encantó al instante, en especial su rama de ciencia de datos, rama en la que espero seguir desarrollándome.

Otra cosa que me apasiona es dibujar y pintar, actividades que estaban dentro de mí pero que se avivaron cuando inició la pandemia en el año 2020.